# Performance of Biometric Quality Measures

Patrick Grother, *Member, IEEE,* and Elham Tabassi, *Member, IEEE*
{pgrother,tabassi}@nist.gov
Image Group, Information Access Division, Information Technology Laboratory
National Institute of Standards and Technology, Gaithersburg, MD, USA

**Abstract**

We document methods for the quantitative evaluation of systems that produce a scalar summary of a biometric sample's quality. We predicate this on the idea that the quality measure predicts performance, whether by design or correlation. We do this abstractly (that is, for arbitrary biometrics) for the general case of a generic black box apparatus that takes a biometric sample as input, and produces some summary output that is intended to be indicative of the expected matching performance from the sample when compared with other samples. We motivate this by reviewing the valuable operational uses of quality values. We detail the quality-performance relationship and give various performance target measures. Finally we present methods and metrics for evaluating a quality algorithm and a procedure for establishing target quality values for a reference biometric data set.

**Index Terms**

evaluation, performance measures, statistical computing, performance evaluation.

## I. Background

BIOMETRIC quality assessment algorithms are increasingly deployed in operational biometric systems [1], and there is recent international consensus in industry [2], academia [3] and government [4] that the quality of a biometric sample is a scalar statement of the predicted recognition performance associated with that sample. The questions then arise of what precisely that association means and how to quantify whether the quality algorithm is actually effective. Prior work in this area and of sample quality generally is limited because it naturally lags matching algorithm development, but also because it is genuinely a new field that is emerging as it is realized that biometric systems fail on certain pathological samples and that tighter and quantitative coupling of the sensor and the feature extraction algorithms is perhaps the primary means of solving this problem.

Our assertion that sample quality measures should be assessed in terms of how well they predict performance is motivated by the fact that performance is ultimately the most relevant goal of a biometric system. We performed quality assessment in large scale offline trials because they offer a robust and repeatable way of evaluating core algorithmic capability. Alonso et al. [5] reviewed five algorithms and used the fingerprints of the multimodal MCYT corpus [6] to compare the distributions of the algorithms' quality assignments, with the result that most of the algorithms behave similarly. We note that finer grained aspects of sample quality can be addressed. For instance Lim et al. [7] trained a fingerprint quality system to predict the accuracy of minutia detection. However, such methods rely on the manual annotation of a data set and this is impractical for all but small datasets, not least because human examiners will disagree in this respect. The virtue of relating quality to performance is that matching trials can be automated and conducted in bulk. We note further that quality algorithms that relate to human perception of a sample quantify performance only as much as the sensitivities of the human visual system are the same as those of a biometric matcher. One additional point is that performance related quality evaluation is agnostic on the underlying technology: it would be improper to force a fingerprint quality algorithm to produce low quality values for an image with few minutia when the target matching algorithm is non-minutia based, as is the case for pattern based methods [8].

We formalize the concept of sample quality as a scalar quantity that is related monotonically to the performance of biometric matchers, under the constraint that at least two samples with their own qualities (as opposed to a pairwise quality) are being compared. We do this in the context of enrollment, verification and identification use-cases. We consider the common and useful case of a quality measure tuned to predict performance of one matcher, and the more difficult case of one that generalizes to other matchers or classes of matchers.

In section II we consider how sample quality is actually used, and this establishes context for the desirable properties of a quality measure that we present in section III. This precedes the main contribution on evaluation in section IV, which discusses the appropriateness of various performance measures as prediction targets for a quality algorithm and then as metrics themselves. In section V we discuss what data should be used for testing a quality algorithm, and document a procedure to construct a reference target database. Conclusions follow in section VI.

The evaluation protocols proposed assume only that the quality algorithm is claimed to predict performance: we do not assume that the algorithm has been standardized nor that its output has any particular distribution. We test the claim by relating quality values to empirical matching results. However, we consider the algorithm to be a black box whose design and intended outputs are determined solely by its author, and we make no assumption of its internal operation.

## II. Uses of Biometric Quality Values

This section describes the roles of a sample quality measure in the various contexts of biometric operations. The quality value here is simply a scalar summary of a sample that is taken to be some indicator of matchability.

### *A. Enrollment Phase Quality Assessment*

Enrollment is usually a supervised process, and it is common to improve the quality of the final stored sample by acquiring as many samples as are needed to satisfy either an automatic quality assessor (the subject of this paper), a human inspector (a kind of quality algorithm), or a matching criterion (by comparison with a second sample acquired during the same session). Our focus on automated systems' needs is warranted regardless of analyses of these other methods, but we do contend that naive human judgment will only be as predictive of a matcher's performance as the human visual system is similar to the matching system's internals, and it is not evident that human and computer matching are functionally comparable. Specifically, human inspectors may underestimate performance on overtly marginal samples. Certainly human inspectors' judgment may be improved if adequate training on the failure modes and sensitivities of the matcher is given to the inspector, but this is often prohibitively expensive or time consuming and not scalable. Immediate matching also might not be predictive of performance over time because same-session samples usually produce unrealistically high match scores. For instance, Fig. 1 shows an example of two same-session fingerprint images that were matched successfully by three commercial vendors despite their obvious poor quality. That said, this paper does not take a position on the merits of doing this. Instead we answer the question that if a quality apparatus is used, is it actually performing?

In any case, by closing this acquire-reacquire loop on a quality measure, a system has a powerful means of improving quality of the enrolled populations' samples. To represent this common use-case we demonstrate, in section IV-B, the effect on performance by enrolling only samples with quality exceeding a threshold.

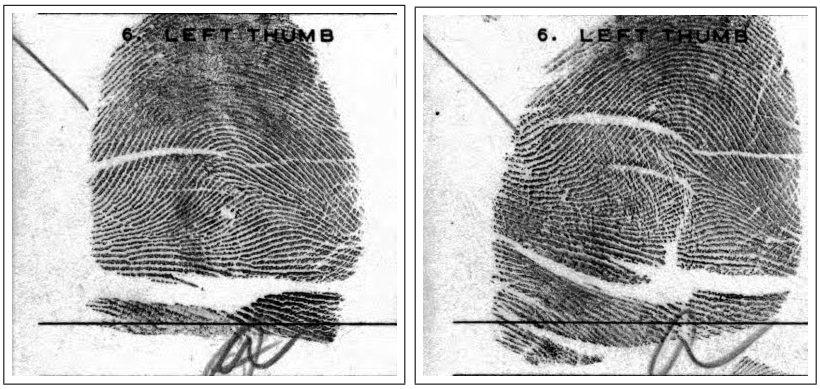
### *B. Quality Assurance Monitoring*

Quality algorithms may also be used as a survey tool. For example when samples are acquired but not immediately matched, recognition problems will not be immediately manifest. Similarly when samples are collected at multiple sites, some disparity may be evident. In both cases, quality values may be aggregated and compared with some historical or geographic baselines. Use of quality values in this role (survey tool) has been documented in [1].

Fig. 7(b) demonstrates the disparity when samples are collected at different sites. It shows the decrease in false non-match rate as poor quality samples are pruned for three fingerprint databases that were collected at different sites using different sensors. Scores from a common verification algorithm was used to compute false non-match rates for all three databases. It is apparent that the green trace has the highest fraction of low quality samples while the blue trace has the lowest.

### *C. Verification Quality Assessment*

During a verification transaction, quality can be improved by closing an acquire-reacquire loop on either a match-score from comparison of new and enrollment samples or on a quality value generated without matching. Indeed




(a) First (b) Second

Fig. 1. Example of same session captures of single finger that despite their poor quality (NFIQ=5) were matched correctly by three leading commercial matchers.

it is common to implement an "up to three attempts" policy in which a positive match is a *de facto* statement that the sample was of good quality - even if the individual happens to be an impostor. Depending on the relative computational expenses of sample matching, reacquisition, and quality measurement, the immediate use of a matcher may not be the best solution.

The key difference here (as compared to the enrollment-phase) is that quality values of both the enrollment and verification samples can be used to predict performance. This two dimensional problem is distinct from the enrollment case where only one quality value is used.

### *D. Identification Quality Assessment*

Quality assessment in identification systems is important for at least three reasons. First, many users often do not have an associated enrollment sample. So a one-to-many match will be an inefficient and inconclusive method of stating whether the authentication sample had high quality. Second, in negative identification systems where users with an enrolled sample are motivated to evade detection, quality assessment can be used to detect and prevent submission of samples likely to perform poorly [9], which may help prevent attempts at spoofing or defeating detection.

Third, identification is a difficult task: it is imperative to minimize both the false non-match rate (FNMR) *and* the false match rate (FMR). To the extent that consistently high quality samples will produce high genuine scores, a high matching threshold can be used and this will collaterally reduce FMR. But in large populations FMR becomes dominant, and this raises the question: can a quality apparatus be trained to be directly predictive of false match likelihood? The word "quality" might be imprecise for such a method, but in any case, the authors can find no publications on metrics that are predictive of FMR.

### *E. Differential Processing*

The quality of a sample can be used to direct or alter the processing of a sample. A number of use-cases fall under this umbrella, and the examples we give here are categorized by the processing stage as follows.

1) *Pre-processing Phase*
   An identification system might apply image restoration algorithms or invoke different feature extraction algorithms for samples with some discernible quality problem. For example, in multi-modal biometrics, the relative qualities of samples of the separate modes may be used to direct, or even elide, a fusion process [10] or trigger the acquisition of second sample if the first preferred mode is of poor quality.
2) *Matching Phase*
   Certain systems may invoke a slower but more powerful matcher when a low-quality samples is input.

3) *Decision Phase*
   The logic that renders acceptance or rejection decisions may depend on the measured quality of the original samples. This would involve a reduction of a verification system's operating threshold for poor quality samples.
4) *Sample Replacement*
   To ameliorate the effects of template ageing, quality (or matching) may be used to determine whether a newly acquired sample should replace the enrolled one. This does not apply to those systems that retain both the old and new samples in support of some multi-instance fusion schemes.
5) *Template Update*
   For the same reasons above, some systems prefer a "soft" combination of features from old and new samples to outright replacement. Such *template update* implementations are somewhat less vulnerable to a false acceptance.

The detailed implementation of such strategies is invariably the responsibility of system integrators, and unlike in the sample re-acquisition role, the use of quality here is usually upon the testers, and/or operators.

## III. Properties of a Quality Measure

This section gives needed background material, including terms, definitions, and data elements, to support quantifying the performance of a quality algorithm.

Throughout this paper we use low quality values to indicate poor sample properties. This is at odds with some systems (for example, the NIST Fingerprint Image Quality (NFIQ) algorithm [11]), for which low values indicate good "quality". Accordingly where we refer to NFIQ in this paper, we transform the values as $Q = 6-\text{NFIQ}$ for consistency with this paper's definition of quality.

### A. Quality as Summary Statistic

Consider a data set $D$ containing two samples, $d_i^{(1)}$ and $d_i^{(2)}$ collected from each of $i = 1, \ldots, N$ individuals. The first sample can be regarded as an enrollment image, the second as a user sample collected later for verification or identification purposes. A discussion of the appropriate composition of this data set for quality algorithm assessment is deferred until section V. For now consider that a quality algorithm $Q$ can be run on the $i$-th enrollment sample to produce a quality value

$$q_i^{(1)} = Q(d_i^{(1)}) \tag{1}$$

and likewise for the authentication (use-phase) sample

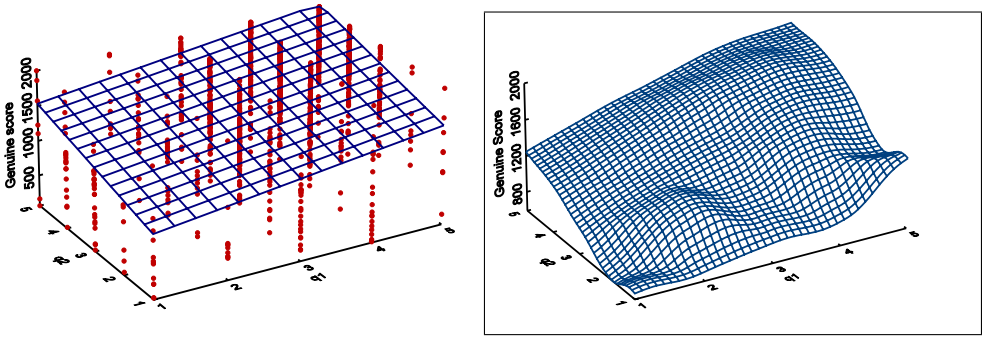
$$q_i^{(2)} = Q(d_i^{(2)}) \tag{2}$$

We have thus far suggested that these qualities are scalars, as opposed to vectors for example. Operationally the requirement for a scalar is not necessary: a vector could be stored and could be used by some predictor. The fact that quality has historically been conceived of as scalar is a widely manifest restriction. For example, BioAPI [12] has a signed single byte value, BioAPI_QUALITY; and the headers of the ISO/IEC biometric data interchange format standards [13] have one or two byte fields for quality. We do not further address the issue of vector quality quantities other than to say that they have been considered (e.g., the defect fields of [2]), and their practical use would require application of a scalar function whose output is actionable and that this should not be an impediment to its use.

### B. Relationship to Matching

We now formalize our premise that biometric quality measures should predict performance. That is, we formalize quality values $q_i$ are related to recognition error rates. A formal statement of such requires an appropriate, relevant and tractable definition of performance. Consider $K$ verification algorithms, $V_k$, that compare pairs of samples (or templates derived from them) to produce match (i.e. genuine) similarity scores

$$s_{ii}^{(k)} = V_k(d_i^{(1)}, d_i^{(2)}) \tag{3}$$




(a) Linear Fit (b) Spline Fit

Fig. 2. Dependence of raw genuine scores on the NFIQ qualities of the two input samples. Note that in this case high quality scores mean better samples.

and similarly non-match (impostor) scores

$$s_{ij}^{(k)} = V_k(d_i^{(1)}, d_j^{(2)}) \quad i \neq j. \tag{4}$$

If we now posit that two quality values can be used to produce an estimate of the genuine similarity score that matcher $k$ would produce on two samples

$$s_{ii}^{(k)} = P(q_i^{(1)}, q_i^{(2)}) + \epsilon_{ii}^{(k)} \tag{5}$$

where the function $P$ is a predictor of a matcher $k$'s similarity scores, and $\epsilon_{ii}$ is the error in doing so for the $i$-th score. Substituting equation (1) gives

$$s_{ii}^{(k)} = P(Q(d_i^{(1)}), Q(d_i^{(2)})) + \epsilon_{ii}^{(k)} \tag{6}$$

and it becomes clear that together $P$ and $Q$ would be perfect imitators of the matcher $V_k$ in equation (3) if the constraint that $Q$ be applied to the samples separately need not apply. This separation is usually a necessary condition for a quality algorithm to be useful because at least half of the time (i.e., enrollment) only one sample is available, see section II. Thus the quality problem is hard first because $Q$ is considered to produce a scalar, and secondly because it is applied separately to the samples. The obvious consequence of this formulation is that it is inevitable that quality values will imprecisely map to similarity scores, i.e., there will be scatter of the known scores, $s_{ii}$, for the known qualities $q_i^{(1)}$ and $q_i^{(2)}$. For example, Fig. 2 shows the raw similarity scores from a commercial fingerprint matcher versus the transformed integer quality scores from the NFIQ algorithm [4], where NFIQ native scores are mapped to $Q = 6-$NFIQ. Fig. 2(a) also includes a least squares linear fit, and Fig. 2(b) shows a spline fit of the same data. Both trend in the correct direction: worse quality gives lower similarity scores. However even though the residuals in the spline fit are smaller than those for the linear, they still are not small. Indeed even with a function of arbitrarily high order, it will not be possible to fit the observed scores perfectly if quality values are discrete (as they are for NFIQ). By including the two fits of the raw data, we do not assert that scores should be linearly related to the two quality values (and certainly not locally cubic). Accordingly we conclude that it is unrealistic to require quality measures to be linear predictors of the similarity scores; instead the scores should be a monotonic function (higher quality samples give higher scores).

Thus our conclusion is that it is futile to consider regresssion methods, because the residuals of equation 5 will not generally have the needed properties for any fit to hold.

### C. Quantized Quality Values

Biometric standards quite reasonably recommend quality values on the range of $[0, 100]$ with the implication that there are that many distinct values (i.e., between 6 and 7 bits). Practically this may not be the case and a coarser quantization, corresponding to $L < 100$ statistically separate levels, is usually achieved. Indeed, although BioAPI [12] states that "no universally accepted definition of quality exists", it goes on to specify four ranges

([0, 25], [26, 50], [51, 75], [76, 100]) with associated meanings: *unacceptable*, *marginal*, *adequate* and *excellent*. This is a tacit acknowledgement that the range $[0, 100]$ is too fine, and that an integer quality value on the range $[1, 4]$ is effectively all that may be needed (or possible). If quality algorithms do not provide 100 statistically distinct levels, an evaluation using $L \ll 100$ would be appropriate. Indeed quantization of a continuous quality metric down to fewer levels may make evaluation easier and/or more robust. For now we avoid the details of the mapping (i.e., from $[1, 100]$ to $[1, L]$) and on whether the tester or the algorithm author should have the responsibility for this, and instead suggest that BioAPI's use of $L = 4$ is a tractible operational definition.

This is *ad hoc*, and clearly a mathematical rationale for $L$ (for example, a criterion against which $L$ can be optimized) is preferable. This could be something like the knees of the distribution functions of the genuine and impostor scores, or $L$ levels based on the separation of the two distributions. An alternative might be to let $L$ be a free parameter in a fitting process, analogous to some discovered intrinsic precision. Regardless of how $L$ is determined, for a quality algorithm to be effective and operationally meaningful, its $L$ quality levels shall be statistically separate.

## IV. Evaluation

This paper's main assertion, that quality should be predictive of performance, has stood so far without a formal specification of how performance should be quantified and whether such performance measures are viable and appropriate. Our overall recommendation is that quality algorithms should be targeted to application-specific performance variables. For verification, these would be the false match and non-match rates. For identification, the metrics would usually be FNMR and FMR [14], but these may be augmented with rank and candidate-list length criteria.

Verification is a positive application, which means samples are captured overtly from users who are motivated to submit high quality samples. For this scenario, the relevant performance metric is the false non-match rate (FNMR) for genuine users because two high quality samples from the same individual should produce a high score. For FMR, it should be remembered that false matches should occur only when samples are biometrically similar (with regard to a matcher). So high quality images should give very low impostor scores, but low quality images should also produce low scores. Indeed it is an undesirable trait for a matching algorithm to produce high impostor scores from low quality samples. (Quality in such cases could be used as a preventive countermeasure. See section II.)

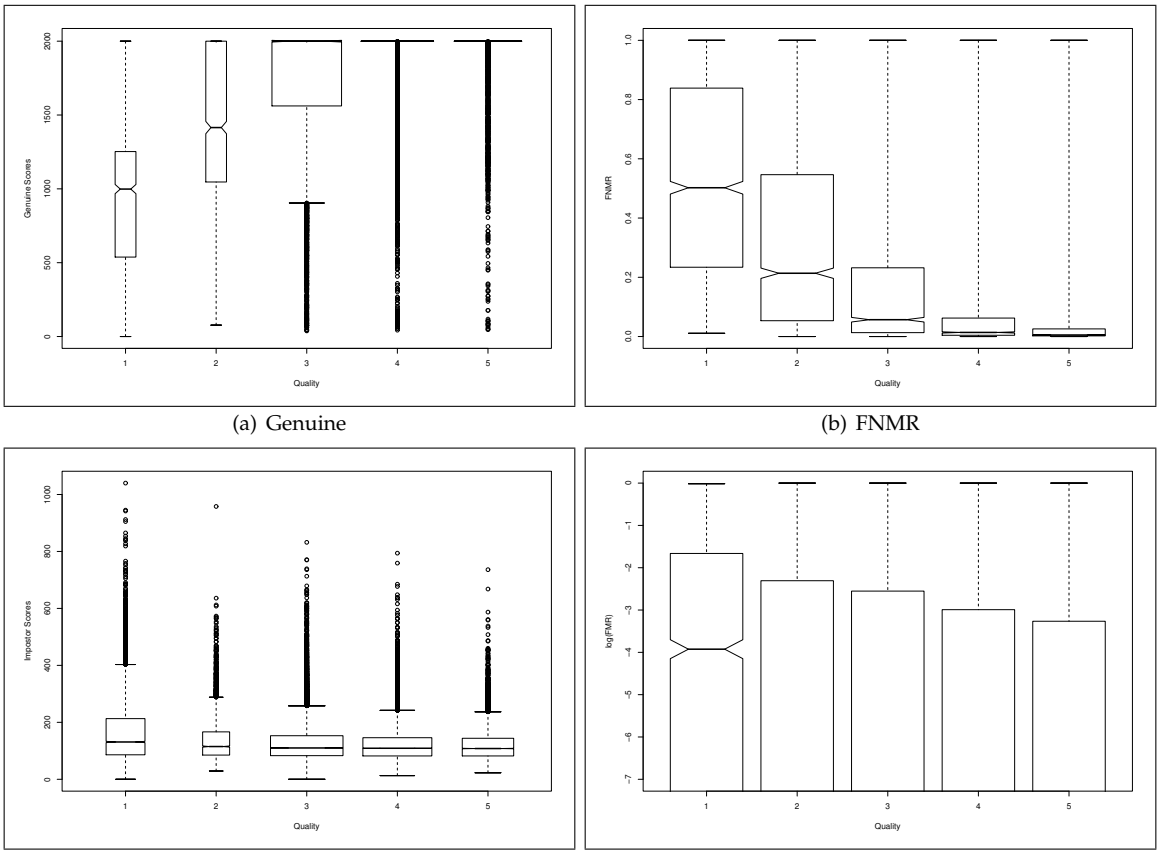
For identification, FNMR is of primary interest. It is the fraction of enrollee searches that do not yield the matching entry on the candidate list. At a fixed threshold, FNMR is usually considered independent of the size of the enrolled population because it is simply dependent on one-to-one genuine scores. However because impostor acceptance, as quantified by FMR, is a major problem in identification systems, it is necessary to ascertain whether low or high quality samples tend to cause false matches.

For a quality algorithm to be effective, an increase in FNMR and FMR is expected as quality degrades. The plots in Fig. 3 show the relationship of transformed NFIQ quality levels to FNMR and FMR. Fig. 3(a) and 3(c) are boxplots of the raw genuine and impostor distribution for each NFIQ quality level of left and right index impressions of 34,800 subjects and scores of a commercial fingerprint matcher. They are accompanied, respectively, by boxplots of FNMR and FMR. The result, that the two error rates decrease as quality improves, is expected and beneficial. The FMR varies less: for quality levels other than the lowest the median FMR values are all zero and are therefore not visible on the log scale. We advocate the use of boxplots as a means of showing that the recognition error rates are statistically separate. If the same quality algorithm had be configured to produce a larger number of levels of quality this separation would at some point disappear. This issue is discussed further in section IV-F.

### A. Combining Two Samples' Quality Values

Biometric matching involves at least two samples. We're faced then with relating performance to quality values $q^{(1)}$ and $q^{(2)}$. This empirical dependence of performance on two often distinct quality values was shown in the plots of Fig. 2. We simplify the analysis by combining the two qualities

$$q_i = H(q_i^{(1)}, q_i^{(2)}) \tag{7}$$




(a) Genuine (b) FNMR

(c) Impostor (d) FMR

Fig. 3. The boxplots show the distribution of genuine scores, FNMR, impostor scores, and FMR for each of five transformed NFIQ quality levels for scores of a commercial matcher. Each quality bin, $q$, contains scores from comparisons of enrollment images with quality $q^{(1)} \geq q$ and subsequent use-phase images with $q^{(2)} = q$, per the discusion in section IV-B.

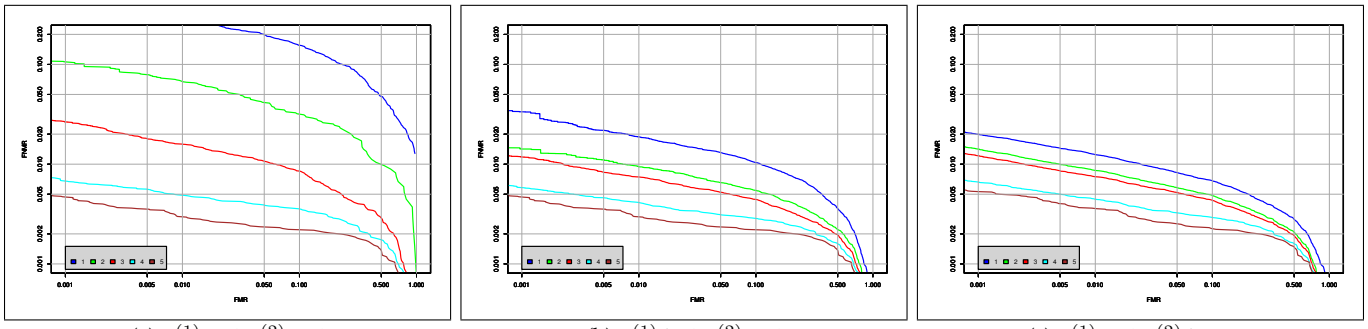
As discussed in section II, quality is primarily useful operationally when a high quality enrollment sample is gathered before a subsequent authentication (use-phase) sample of less controlled quality. To capture this concept we consider $H(x, y) = \min(x, y)$ i.e. the worse of two samples drives the similarity score. We also consider the arithmetric and geometric means, $H(x, y) = (x + y)/2$ and $H(x, y) = \sqrt{xy}$ (see [15]), and the difference function $H(x, y) = |x - y|$ to investigate dependence of similarity score on samples of different quality. We note that whatever $H$ is used it should be well defined for allowed values of $x$ and $y$ (e.g., positive values for the geometric mean).

We now describe four methods for the evaluation of quality. All four consider the use of combination functions, $H$, which are specifically compared in section IV-C.

### *B. Rank-ordered DETs*

A quality algorithm is useful if it can at least give an ordered indication of an eventual performance. For example for $L$ discrete quality levels there should notionally be $L$ DET characteristics[1]. In the studies [3], [15], and [11] that have evaluated quality, DET's are the primary choice. We recognize that DET's are widely understood, even expected, but note three problems with their use: they confound genuine and impostor scores; they are used without a test of the significance of the separation of $L$ levels; and partitioning of the data for their computation is underreported or non-standard.

[1]The DET, which plots FNMR vs. FMR on log scales, is ubiquitously used to summarize verification performance. It conveys the same information as the receiver operating characteristic which plots $1-$FNMR on a linear scale.



(a) $q^{(1)} = i, q^{(2)} = i$ (b) $q^{(1)} \geq i, q^{(2)} = i$ (c) $q^{(1)} = i, q^{(2)} \geq -\infty$

Fig. 4. Quality ranked DET characteristics. Each plot shows five traces corresponding to five transformed NFIQ levels.

On the partitioning of the data, we give three methods for the quality-ranked DET computation. All three use $N$ paired matching images with integer qualities $q_i^{(1)}$ and $q_i^{(2)}$ on the range $[1, L]$. Associated with these are $N$ genuine similarity scores, $s_{ii}$, and up to $N(N-1)$ impostor scores, $s_{ij}$ where $i \neq j$, obtained from some matching algorithm. All three methods compute a DET characteristic for each quality level $k$. For all thresholds $s$, the DET is a plot of FNMR($s$) = $M(s)$ versus FMR($s$) = $1 - N(s)$, where the empirical cumulative distribution functions $M(s)$ and $N(s)$ are computed, respectively, from sets of genuine and impostor scores. The three methods of partitioning differ in the contents of these two sets. The simplest case uses scores obtained by comparing authentication and enrollment samples whose qualities are both $k$. This procedure (see for example, [16]) is common but overly simplistic. By plotting

$$\begin{aligned} \text{FNMR}(s,k) &= \frac{\left|\left\{s_{ii}:\ s_{ii} \leq s,\quad q_i^{(1)}=q_i^{(2)}=k\right\}\right|}{\left|\left\{s_{ii}:\ s_{ii} \leq \infty,\ q_i^{(1)}=q_i^{(2)}=k\right\}\right|} \\ \text{FMR}(s,k) &= \frac{\left|\left\{s_{ij}:\ s_{ij} > s,\quad q_i^{(1)}=q_j^{(2)}=k,\ i \neq j\right\}\right|}{\left|\left\{s_{ij}:\ s_{ij} > -\infty,\ q_i^{(1)}=q_j^{(2)}=k,\ i \neq j\right\}\right|} \end{aligned} \tag{8}$$

the DETs for each quality level can be compared. Although for a good quality algorithm this will show an ordered relationship between quality and error rates, it is not representative of an operational use of quality. Rather by computing performance from scores obtained by comparing authentication samples of quality $k$ with enrollment samples of quality greater than or equal to $k$,

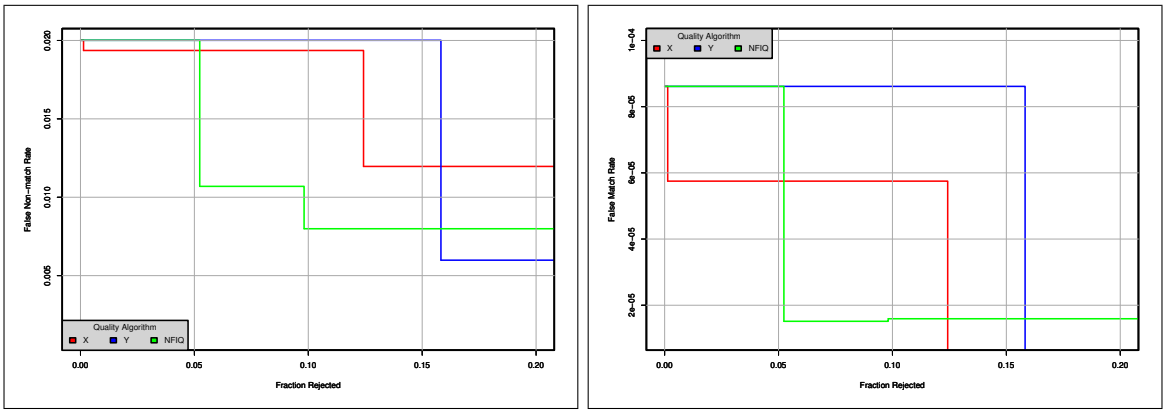
$$\begin{aligned} \text{FNMR}(s,k) &= \frac{\left|\left\{s_{ii}:\ s_{ii} \leq s,\quad q_i^{(1)} \geq k,\ q_i^{(2)}=k\right\}\right|}{\left|\left\{s_{ii}:\ s_{ii} \leq \infty,\ q_i^{(1)} \geq k,\ q_i^{(2)}=k\right\}\right|} \\ \text{FMR}(s,k) &= \frac{\left|\left\{s_{ij}:\ s_{ij} > s,\quad q_i^{(1)} \geq k,\ q_j^{(2)}=k,\ i \neq j\right\}\right|}{\left|\left\{s_{ij}:\ s_{ij} > -\infty,\ q_i^{(1)} \geq k,\ q_j^{(2)}=k,\ i \neq j\right\}\right|} \end{aligned} \tag{9}$$

we model the situation in which the enrollment samples are at least as good as the authentication (i.e., user submitted) samples. Such a use of quality would lead to failures to acquire for the low quality levels.

If instead we compare performance across *all* authentication samples against enrollment samples of quality greater than or equal to $k$,

$$\begin{aligned} \text{FNMR}(s,k) &= \frac{\left|\left\{s_{ii}:\ s_{ii} \leq s,\quad q_i^{(1)} \geq k\right\}\right|}{\left|\left\{s_{ii}:\ s_{ii} \leq \infty,\ q_i^{(1)} \geq k\right\}\right|} \\ \text{FMR}(s,k) &= \frac{\left|\left\{s_{ij}:\ s_{ij} > s,\quad q_i^{(1)} \geq k,\ i \neq j\right\}\right|}{\left|\left\{s_{ij}:\ s_{ij} > -\infty,\ q_i^{(1)} \geq k,\ i \neq j\right\}\right|} \end{aligned} \tag{10}$$

we model the situation where quality control is applied only during enrollment. If repeated enrollment attempts fail to produce a sample with quality above some threshold, a failure-to-enroll (FTE) would be declared. This scenario is common and possible because enrollment, as an attended activity, tends to produce samples of better quality than authentication.




(a) Finger - FNMR (b) Finger - FMR

Fig. 5. Error vs. reject performance for three fingerprint quality methods. Figs. (a) and (b) show reduction in FNMR and FMR at a fixed threshold as up to 20 % of the low quality samples are rejected. The similarity scores come from a commercial matcher.

The considerable differences between these three formulations are evident in the DETs of Fig. 4 for which the NFIQ algorithm [4] for the predicting performance of a commercial fingerprint system was applied to over 61,993 genuine and 121,997 impostor comparisons (NFIQ native scores were transformed to $Q = 6-$NFIQ). In all cases the ranked separation of the DETs is excellent across all operating points. We recommend that equation (9), as shown in Fig. 4(b), be used because of it is a more realistic operational model.

However, as relevant as DET curves are to expected performance, we revisit here a very important complication. Because DET characteristics quantify the separation of the genuine and impostor distributions and combine the effect of quality on both genuine and impostor performance, we lose sight of the separate effects of quality on FNMR and FMR.

That quality should be evaluated at all in relation to impostor performance (i.e. FMR) is dubious. For example, does a biometric recognition system produce a low impostor score when the two samples are of low quality? Perhaps, but does it also produce lower impostor scores when the samples are of high quality? Under what circumstances are the impostor scores high? (Such questions may be simpler to answer for a fingerprint quality apparatus that predicts a minutiae based matcher's performance on the basis of number and type, etc. of minutia.) In any case, we conclude that DETs, while familiar and highly relevant, confound genuine and impostor scores. As an alternative, we propose the boxplots of Fig. 3 and the methods advanced in the next section.

### *C. Error vs. Reject Curves*

An alternative approach to the DET curve is the error-vs-reject curve. Again this models the operational case in which a quality apparatus is used to reject low quality samples for the purpose of improving performance. Consider that a pair of samples (from the same subject), with qualities $q_i^{(1)}$ and $q_i^{(2)}$, are compared to produce a score $s_{ii}^{(k)}$, and this is repeated for $N$ such pairs.

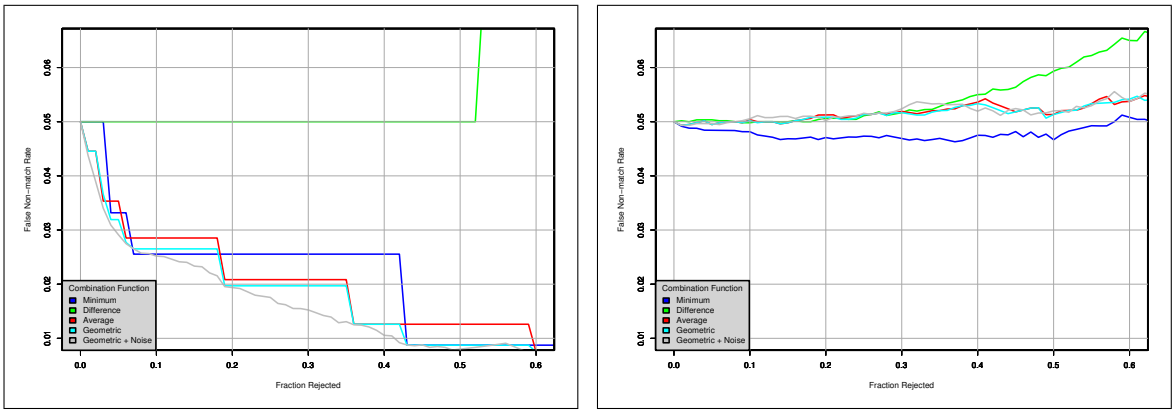
Applying equation 7, we form the set of low quality entries

$$R(u,v) = \left\{ j \ : \ q_j^{(1)} < u, \quad q_j^{(2)} < v \right\} \tag{11}$$

and use it to exclude the associated similarity scores from the computation of "high quality" FNMR at some fixed threshold $t$:

$$\text{FNMR}(t,u,v) = \frac{|\{s_{jj} : s_{jj} \leq t, j \notin R(u,v)\}|}{|\{s_{jj} : s_{jj} \leq \infty\}|} \tag{12}$$

and then plot this for a threshold $t$ corresponding to some reasonable false non-match rate of interest, $f$. The threshold is obtained from the quantile function of the empirical cumulative distribution function of the scores, $t = M^{-1}(1-f)$.




(a) Finger (b) Face

Fig. 6. Dependence of the error vs. reject characteristic on the quality combination function H(.). The plots show, for a fixed threshold, the decrease in FNMR as up to 60% of the low NFIQ quality values are rejected. The similarity scores come from a commercial matcher.

The analagous equations for the one-dimensional case (i.e., min(), etc.) of section IV-A are:

$$R(u) = \left\{ j \ : \ H(q_j^{(1)}, q_j^{(2)}) < u \right\} \tag{13}$$

and then state false non-match performance as the proportion of non-excluded scores below the threshold.

$$\text{FNMR}(t, u) = \frac{|\{s_{jj} : s_{jj} \leq t, j \notin R(u)\}|}{|\{s_{jj} : s_{jj} \leq \infty\}|} \tag{14}$$

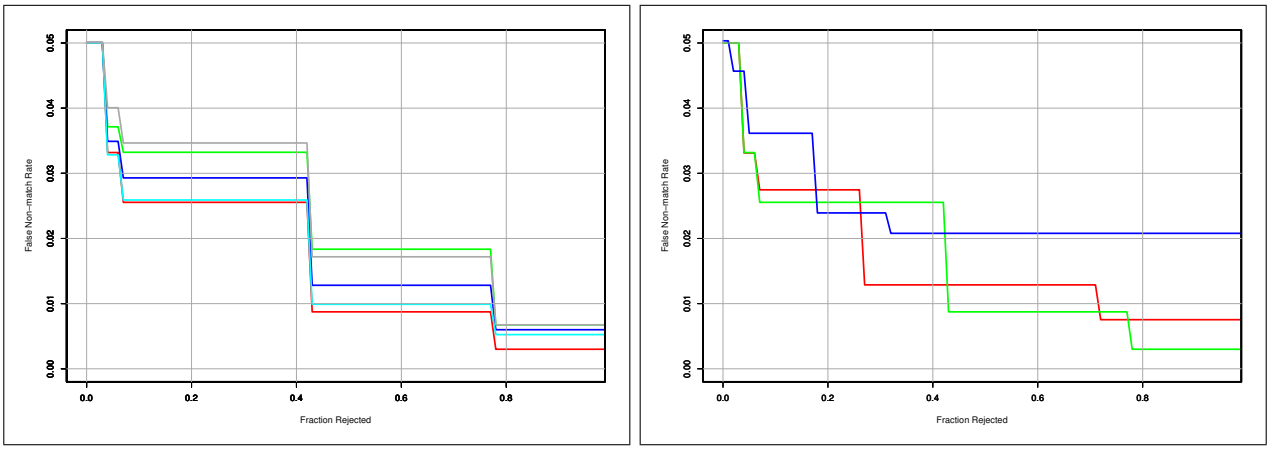
If the quality values are perfectly correlated with the genuine scores, then when we set $t$ to give an overall FNMR of $x$ and then reject proportion $x$ with the lowest qualities. A recomputation of FNMR should be zero. Thus, a good quality metric correctly labels those samples that cause low genuine scores as poor quality. For a good quality algorithm, FNMR should decrease quickly with the fraction rejected. The results of applying this analysis are shown in Fig. 5. Note that the curves for each of the three fingerprint quality algorithms trend in the correct direction, but that the even after rejection of 20 % the FNMR value has fallen only by about a half from its starting point. Rejection of 20 % is probably not an operational possibility unless an immediate reacquisition can yield better quality values for those persons. Note, however, that for NFIQ, the improvement is achieved after rejection of just 5 %. We suggest that correct (i.e. rapid) rejection should be the central challenge to quality algorithm designers.

Figure 6 shows error vs. reject behavior for the NFIQ quality method when the various $H(q_1, q_2)$ combination functions of section IV-A are used. Between the minimum, mean, and geometric mean functions there is little difference. The geometric mean is best (absent a significance test) with steps occuring at values corresponding to the square roots of the product of NFIQ values (i.e. $1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 16, 20, 25$). The gray line in the figure shows $H = \sqrt{q_1 q_2} + N(0, 0.01)$ where the gaussian noise serves to randomly reject samples within a quality level and produces an approximation of the lower convex hull of the geometric mean curve. The green line result, for $H = |q_1 - q_2|$, shows that transformed genuine comparison score is unrelated to the difference in the qualities of the samples. Instead the conclusion is that FNMR is related to monotonic functions of the two values. The generality of this result to other quality methods is not known.

### *D. Generalization to Multiple Matchers*

It is a common contention that the efficacy of a quality algorithm is necessarily tied to a particular matcher. We observe that this one-matcher case is commonplace and useful in a limited fashion and should therefore be subject to evaluation. However, we also observe that it is possible for a quality algorithm to be capable of generalizing across *all* (or a class of) matchers, and this too should be evaluated.

Generality to multiple matchers can be thought of as an interoperability issue: can supplier A's quality measure be used with supplier B's matcher? Such a capability will exist to the extent that pathological samples do present




(a) Across Matchers (b) Across Datasets

Fig. 7. Error vs. reject characteristics showing how NFIQ generalizes across (a) five verification algorithms and (b) three operational data sets. The steps in (a) occur at the same rejection values because the matchers were run on a common database.

problems to both A and B's matching algorithms. However, the desirable property of generality exposes another problem: we cannot expect performance to be predicted absolutely because there are good and bad matching systems. A system here includes all of the needed image analysis and comparison tasks. Rather we assert that a quality algorithm intended to predict performance generally need only be capable of giving a relative or rank ordering i.e. low quality samples should give lower performance than high quality samples.

The plots of Fig. 7 quantify this generalization for the NFIQ system using the error vs. reject curves of section IV-C. Fig. 7(a) includes five traces, one for each of five verification algorithms. Large vertical spread of the traces indicates variation in how well NFIQ is functioning for various matchers, but nonetheless shows its generalization to other matchers.

### *E. Number of Levels of Quality*

A quality metric is more useful if operationally it may be thresholded at one of many distinct operating points. Thus a discrete-valued quality measure is better if performance is significantly different for level $q_k$ than for $q_{k-1}$ for all levels $1 \leq k \leq K$. Having already stated that FNMR should be monotonic in the quality value, $\text{FNMR}_k \leq \text{FNMR}_{k-1}$, we now additionally require that the quality levels are statistically distinct. If they are not, they could be mapped to fewer levels that are statistically distinct. Real values can be quantized. Formally we propose to test this by using the Kolmogorov Smirnov (KS) test to determine whether the distribution of the genuine scores for level $q_k$ is distinct from that of $q_{k-1}$. The KS test is non-parametric, distribution-free and simple. The KS statistic is simply the maximum absolute difference between the two distributions' cumulative distributions functions.

Table I shows example results for two fingerprint quality methods. In both cases the observed KS statistic values are smaller for higher quality levels (where performance is always very high) and are significant: The p-value exceeds $10^{-7}$ on only one occasion. A higher p-value there would have indicated that quality method Y's level 5 and 6 are insignificantly different. The results do not demonstrate such behavior presumably because the algorithms were created with a reasonable number of levels as a design parameter.

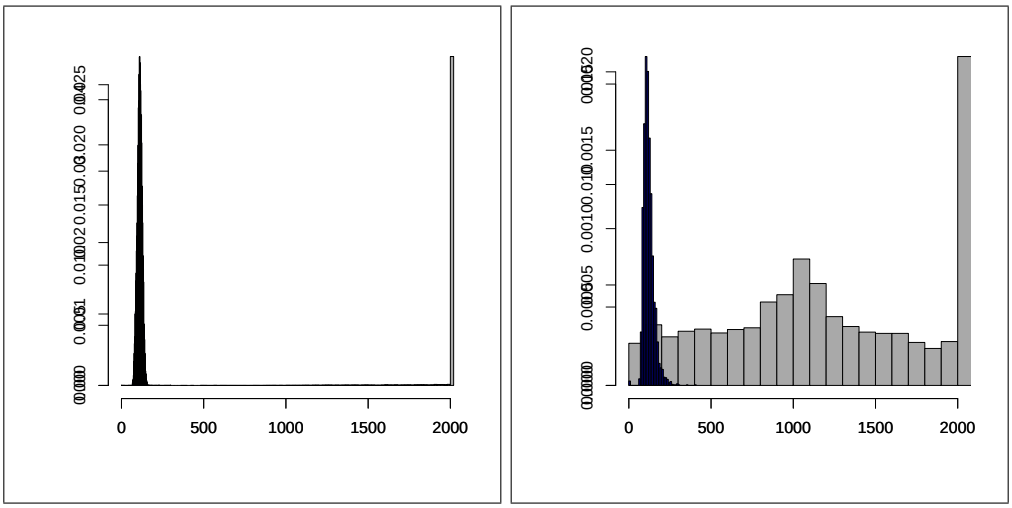
### *F. Measuring Separation of Genuine and Impostor Distributions*

We can evaluate quality algorithms on their ability to predict how far a genuine score will lie from its impostor distribution. This means instead of evaluating a quality algorithm solely based on its FNMR (i.e., genuine score distribution) prediction performance, we can augment the evaluation by including a measure of FMR because correct identification of an enrolled user depends both on correctly finding the match and on rejecting the non-matches. Note also that a quality algorithm could invoke a matcher to compare the input sample with some internal background samples to compute sample mean and standard deviation.

| Quality | $q$ = | No. Scores at Level | | KS Statistic Between Genuine Distribution of | p |
|---|---|---|---|---|---|
| Method | $min(q^{(1)}, q^{(2)})$ | $k-1$ | $k$ | $\{s_{ii} : q_i^{(1)} = q_i^{(2)} = q\}$ and | value |
| | | | | $\{s_{ii} : q_i^{(1)} = q_i^{(2)} = q-1\}$ | |
| X | 2 | 3647 | 3191 | 0.20 | 0 |
| X | 3 | 3191 | 20553 | 0.34 | 0 |
| X | 4 | 20553 | 24297 | 0.24 | 0 |
| X | 5 | 24297 | 17975 | 0.08 | 0 |
| Y | 2 | 11023 | 4878 | 0.30 | 0 |
| Y | 3 | 4878 | 6637 | 0.05 | 0 |
| Y | 4 | 6637 | 8440 | 0.07 | 0 |
| Y | 5 | 8440 | 10751 | 0.05 | 0 |
| Y | 6 | 10751 | 11902 | 0.02 | 0.006 |

TABLE I

KS TEST FOR SEPARATION OF QUALITY-SPECIFIC GENUINE SCORE DISTRIBUTIONS. THE DATA APPLY TO 69663 GENUINE FINGERPRINT COMPARISONS.




(a) Best NFIQ (b) Poorest NFIQ

Fig. 8. There is a higher degree of separation between the genuine and impostor distribution for better quality samples as measured by NFIQ.

Fig. 8(a) and 8(b) show, respectively, the genuine and impostor distributions for the best and worst NFIQ levels. The overlapping of genuine and impostor distributions for the poorest NFIQ means higher recognition errors for that NFIQ level, and vice versa; the almost complete separation of the two distribution for the best quality samples indicates lower recognition error. NFIQ was trained to specifically exhibit this behavior.

We again consider the KS statistic. For better quality samples, a larger KS test statistic (i.e. higher separation between genuine and impostor distribution) is expected. Each row of Table II shows KS statistics for one of the three quality algorithms that we tested. KS statistics for each quality levels $u = 1, \ldots, 5$ are computed by first computing the genuine (i.e., $\{s_{ii} : (i,i) \in R(u)\}$) and impostor (i.e., $\{s_{ij} : (i,j) \in R(u), i \neq \text{J}\}$) empirical cumulative distributions, where $R(u) = \{(i,j) : H(q_i^{(1)}, q_j^{(2)}) = u\}$. Thereafter the largest absolute difference between the genuine and impostor distributions of quality $u$ is measured and plotted. (Note that to keep quality algorithm providers anonymous we only reported KS statistics of the lowest four quality levels.)

## V. QUALITY REFERENCE DATA SETS

This section addresses two issues: what data should be used for testing a quality apparatus and how to construct a reference quality annotated database to support evaluation. We precede those topics by critiquing, in the next subsection, the use of deliberatly degraded sample data.

TABLE II
KS STATISTICS FOR QUALITY LEVELS OF THREE QUALITY ALGORITHMS

| **Quality Algorithm** | $Q = 1$ | $Q = 2$ | $Q = 3$ | $Q = 4$ |
|---|---|---|---|---|
| Quality Algorithm 1 | 0.649 | 0.970 | 0.988 | 0.993 |
| Quality Algorithm 2 | 0.959 | 0.995 | 0.996 | 0.997 |
| Quality Algorithm 3 | 0.918 | 0.981 | 0.994 | 0.997 |

We note that the use of "laboratory" data, i.e. deliberately collected low quality samples (e.g., by deliberate non-conformance to a data acquisition standard) should be deprecated for reasons of operational relevance. Rather, a quality metric should be evaluated on operationally relevant data.

### *A. Data to be used for testing*

Samples that have been deliberately collected with specific defects, could be used for testing. For example, quality could be degraded by misfocusing the camera. Such data have several notable uses: development of a quality measurement apparatus, teaching best practice by counterexample, and assessing the performance of a product intended to test the conformance of an image or signal to an underlying standard[2]. However, we argue that this type of data should not be used for evaluation for four reasons. First, such data is by definition laboratory data and therefore would lack application-specific operational relevance. Second, it embeds assumptions about the performance sensitivites of matching algorithms. Third, it would be difficult or impossible to collect samples that express all possible combinations of quality defects and particularly with their natural frequency and to their natural degree. Finally, the laboratory data would not ordinarily be available in large quantities.

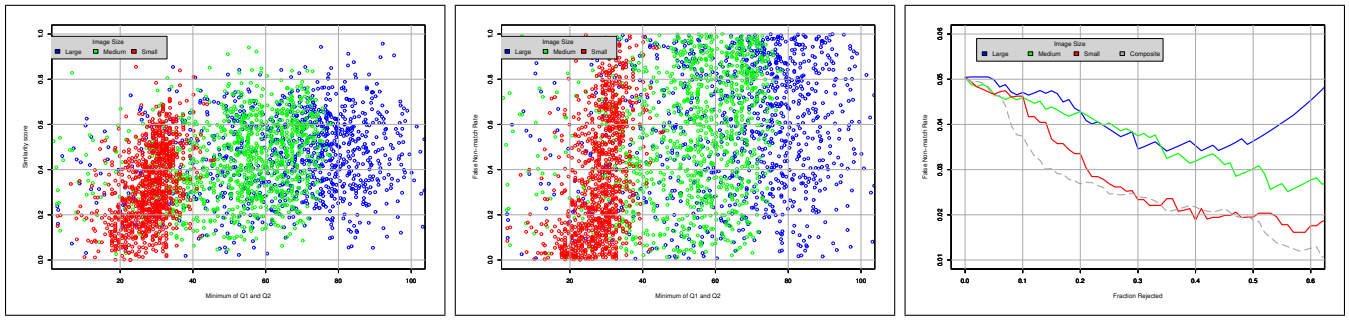
Instead this paper considers the use of operationally representative data, i.e. samples harvested during real-world usage or from a relevant scenario test [17]. By definition, this has the advantage of having relevance to the operation. We showed examples of such data in section IV-C.

However, if a test compares quality algorithms or is making a more general assessment of the technology, then an aggregated corpus that spans the quality spectrum might be employed. Such a set might include fingerprint images gathered from employees during an access control enrollment and subsequently authentication and also samples collected outdoors and from persons detained in adverse law enforcement environments. This construction, unlike the dedicated laboratory collection described above, does not manipulate the sample acquistion process.

To illustrate the importance of using an aggregated corpus for evaluation, we use the frontal *fa* and *fb* images from the Color FERET database [18] at full, half and quarter resolutions. We expose these to a quality algorithm and matcher from the same supplier. The reduction in image size forcibly induces the reductions in both quality and match scores evident in Fig. 9. Note, however, that for any one of the three point clouds in Fig. 9(a), there is large variation in score in relation to quality - a trend that is not ameliorated by plotting $M(s)$ instead (Fig. 9(b)). This reflects the difficulty of the face quality problem.

The final graph, Fig. 9(c), shows the error versus reject performance for each of the image sizes separately as well as in grey for the aggregate data set. That the composite curve is lower than the others demonstrates some probitive value in using composite sets. Also worthy of note is that the performance of the quality algorithm at any of the three sizes is superior to that of the same algorithm used in Fig. 6(b) which was computed using images from a different corpus. Those images, while about the same size as the half-size FERET images, are more highly compressed. The conclusion is that the more homogenous the corpus, the less well a quality algorithm will perform. We should emphasize that the algorithm was provided to the authors without any claim of efficacy or recommended domain of use.

[2]For example, the ISO/IEC 19794-5 Face Recognition Interchange Format standard puts quantitative limits on the amount of quality related degradation from such as blur, non-frontal pose, and the number of grey levels.




(a) Score vs. $\min(q_1, q_2)$ (b) FNMR vs. Min Quality (c) Error vs. Reject

Fig. 9. Score values, FNMR values and error vs. reject curves, for a face quality metric applied to a composite database.

### *B. Construction of a Reference Data Set*

In this section, we define a procedure for constructing a performance-oriented reference database. The result is a set of samples annotated with a target quality value. The value is essentially a consensus similarity score classificiation from a set of matchers. Such reference sets are of primary use to quality algorithm developers working on the general problem. The same method would be of use to tune a quality algorithm to an operational situation in which the matcher and kind of data are known and available.

The input to our procedure is a representative sample database. The output is an annotation of each sample with a scalar quality target. The method presumes the availability of a representative matching algorithm, which will be used to compare samples to produce both genuine and impostor similarity scores. It is therefore implied that two or more samples per person are available.

*1) Data:* Data gathered in a target operational application would be most realistic. Contemporary matchers perform extremely well on most images, and it is therefore necessary to preferentially stack the reference set with samples that are naturally problematic to the matcher. For example, for a reference fingerprint data set to span the quality spectrum, it should be balanced in terms of finger position (right/left index/thumb/middle), finger impression (roll/plain/flat), sex, age, and capture device. Lack of data often renders it difficult to create such a balanced dataset.

*2) Target Quality Assignment :* We seek to assign a ground-truth quality score to each image in a reference dataset. We ensure that the quality values are representative of performance by associating the image with similarity scores as follows. Consider a biometric corpus containing two samples, $d_i^{(1)}$ and $d_i^{(2)}$, for each of $N$ individuals, $i = 1, \ldots, N$. The first samples represent enrollment samples, and the second samples represent those for authentication. The following procedure assigns quality values $q_i^{(1)}$ and $q_i^{(2)}$ to all images in the corpus.

For each person $i$:

1) Compare the first and second samples using the $k$-th matcher to produce genuine score. Repeating equation 3:
$$s_{ii}^{(k)} = V_k(d_i^{(1)}, d_i^{(2)}) \tag{15}$$
2) Use the $k$-th matcher to compare the first sample from person $i$ with the second sample from all $j = 1, \ldots, N$ and $i \neq j$ other persons. The result is $J = N - 1$ impostor scores:
$$s_{ij}^{(k)} = V_k(d_i^{(1)}, d_j^{(2)}) \tag{16}$$
(This is essentially eq. 4.)
3) Insert $i$ into set $\mathcal{T}$ if its genuine score is larger than all its impostor scores, i.e. $s_{ii}^{(k)} > s_{ij}^{(k)} \ \forall j$. This is a rank 1 condition.
4) For the first sample of each person $d_i^{(1)}$, compute the sample mean and standard deviation of its $J$ associated impostor scores

TABLE III
BINNING NORMALIZED MATCH SCORE

| Bin | Range of normalized match score |
|---|---|
| 1 | $\{z_i : -\infty \leq z_i < C^{-1}(0)\}$ |
| 2 | $\{z_i : C^{-1}(0) \leq z_i < W^{-1}(1)\}$ |
| 3 | $\{z_i : W^{-1}(1) \leq z_i < C^{-1}(0.25)\}$ |
| 4 | $\{z_i : C^{-1}(0.25) \leq z_i < C^{-1}(0.75)\}$ |
| 5 | $\{z_i : C^{-1}(0.75) \leq z_i\}$ |

$$m_i = J^{-1} \sum_{j=1}^{J} s_{ij}^{(k)} \tag{17}$$

$$\sigma_i = (J-1)^{-1} \sum_{j=1}^{J} \left( s_{ij}^{(k)} - m_i \right)^2 \tag{18}$$

5) Normalize the genuine score from eq. 15 using the impostor statistics

$$z_i = (s_{ii} - m_i)/\sigma_i \tag{19}$$

Once all normalized similarity scores have been computed:

1) Compute two empirical cumulative distribution functions: One for the top-ranked genuine scores of set $\mathcal{T}$

$$C(z) = \frac{|\{z_i : i \in \mathcal{T}, z_i \leq z\}|}{|\{z_i : i \in \mathcal{T}, z_i \leq \infty\}|} \tag{20}$$

and another for those not in that set.

$$W(z) = \frac{|\{z_i : i \notin \mathcal{T}, z_i \leq z\}|}{|\{z_i : i \notin \mathcal{T}, z_i \leq \infty\}|} \tag{21}$$

These cumulative distribution functions are plotted in Fig. 10 for live-scan images of the right-index fingers of 6000 individuals and scores of a commercial fingerprint matcher.

2) Bin normalized match score range into $K$ bins based on quantiles of the normalized match score distribution. One strategy, for $K = 5$, is shown in Table III in which $F^{-1}$ is the quantile function, and $F^{-1}(0)$ and $F^{-1}(1)$ denote the empirical minima and maxima, respectively. If $W^{-1}(1) \geq C^{-1}(0.25)$ an appropriate quartile of $C(z)$ must be selected.
3) Sample $d_i$ is assigned target quality $q_i$ corresponding to the bin of its normalized match score $z_i$ from eq. (19).
4) The procedure is repeated for sample $d_i^{(2)}$ by swapping indices 1 and 2 in equations 15 and 16. Since one sample will have an imposter distribution different from another, two different samples of the same subject may have different normalized match scores and therefore different quality values.
5) The procedure is repeated for scores of all $V$ matchers.
6) Samples with identical quality assignments from *all* $V$ matchers become members of the Quality Reference Dataset. Those without unanimity can are discarded.

This procedure has been used to form NFIQ training and compliance set [19], only with different bin boundaries. The NFIQ boundaries shown in Table IV were set by visual inspection to give useful categorization of the normalized match score statistic.

## VI. CONCLUSION

Biometric quality assessment is an operationally important and difficult problem that is nevertheless massively under-researched in comparison to the primary feature extraction and pattern recognition tasks. In this paper, we enumerated the ways in which it is useful to compute a quality value from a sample. In all cases the ultimate intention is to improve matching performance. We asserted therefore that quality algorithms should be developed

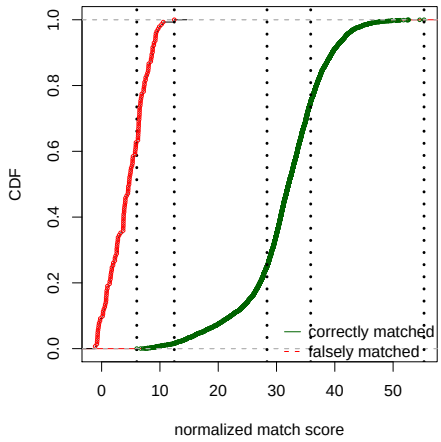



Fig. 10. Empirical cumulative distribution functions for the top-ranked genuine scores and for the imposter scores. The vertical line are one possible way of binning normalized match score. Samples are assigned quality numbers corresponding to the bin of their normalized match score.

TABLE IV
BIN BOUNDARIES (OR RANGE OF NORMALIZED MATCH SCORE) FOR NFIQ

| NFIQ | Label | Range |
|---|---|---|
| 1 | poor | $\{z_i : -\infty \leq z_i < W^{-1}(0.75)\}$ |
| 2 | fair | $\{z_i : W^{-1}(0.75) \leq z_i < C^{-1}(0.05)\}$ |
| 3 | good | $\{z_i : C^{-1}(0.05) \leq z_i < C^{-1}(0.2)\}$ |
| 4 | very good | $\{z_i : C^{-1}(0.2) \leq z_i < C^{-1}(0.6)\}$ |
| 5 | excellent | $\{z_i : C^{-1}(0.6) \leq z_i < C^{-1}(1)\}$ |

to explicitly target matching error rates, and not human perceptions of sample quality. To this end, we defined a procedure for the annotation of a reference sample set with target quality values. We gave several means for assessing the efficacy of quality algorithms. We reviewed the existing practice, cautioned against the use of detection error tradeoff characteristics as the primary metrics, and instead advanced boxplots and error vs. reject curves as preferable. We suggest that algorithm designers should target false non-match rate as the primary performance indicator.

In conclusion, we posit that quality summarization as a predictor of recognition performance is a difficult problem, and we encourage the academic community to consider the problem and extend the quantitative methods of this paper in advancing their work.

## REFERENCES

[1] T. Ko and R. Krishnan, "Monitoring and reporting of fingerprint image quality and match accuracy for a large user application," in *Proceedings of the 33-rd Applied Image Pattern Recognition Workshop*. IEEE Computer Society, 2004, pp. 159–164.
[2] D. Benini et al., *ISO/IEC Report of the Ad Hoc Group on Biometric Quality: Document N1128*, 3rd ed., JTC1 / SC37 / Working Group 3, May 2005, http://isotc.iso.org/isotcportal.
[3] Y. Chen, S. Dass, and A. Jain, "Fingerprint quality indices for predicting authentication performance," in *Procedings of the Audio- and Video-based Biometric Person Authentication (AVBPA)*, July 2005, pp. 160–170.
[4] E. Tabassi, *Fingerprint Image Quality, NFIQ*, NISTIR 7151 ed., National Institute of Standards and Technology, 2004.
[5] F. Alonso-Fernandez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "A review of schemes for fingerprint image quality computation," in *COST 275 - Biometrics-based recognition of people over the internet*, October 2005.
[6] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro, "Mcyt baseline corpus: a bimodal biometric database," *Proceedings of the IEE Conference on VISP*, vol. 150, no. 6, pp. 395–401, December 2003.
[7] E. Lim, X. Jiang, and W. Yau, "Fingerprint quality and validity analysis," in *Proceedings of the IEEE Conference on Image Processing*, vol. 1, September 2002, pp. 469–472.

[8] Bioscrypt Inc., *Systems and Methods with Identify Verification by Comparison and Interpretation of Skin Patterns such as Fingerprints*, June 1999, http://www.bioscrypt.com.
[9] L. M. Wein and M. Baveja, "Using fingerprint image quality to improve the identification performance of the u.s. visit program," in *Proceedings of the National Academy of sciences*, 2005, www.pnas.org/cgi/doi/10.1073/pnas.0407496102.
[10] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Discriminative multimodal biometric authentication based on quality measures," *Pattern Recognition*, vol. 38, no. 5, pp. 777–779, May 2005.
[11] E. Tabassi, "Fingerprint image quality, nfiq," in *IEEE International Conference on Image Processing ICIP-05*, Genoa, Italy, September 2005.
[12] C. Tilton et al., *The BioAPI Specification*, American National Standards Institute, Inc., 2002.
[13] ISO/IEC JTC1 / SC37 / Working Group 3, "ISO/IEC 19794 Biometric Data Interchange Formats," 2005, http://isotc.iso.org/isotcportal.
[14] A. J. Mansfield, *ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework*, FDIS ed., JTC1 / SC37 / Working Group 5, August 2005, http://isotc.iso.org/isotcportal.
[15] J. Fierrez-Aguilar, L. Muñoz-Serrano, F. Alonso-Fernandez, and J. Ortega-Garcia, "On the effects of image quality degradation on minutiae and ridge-based automatic fingerprint recognition," in *IEEE International Carnahan Conference on Security Technology*, October 2005.
[16] D. Simon-Zorita, J. Ortega-Garcia, J. Fierrez-Aguilar, and J. Gonzalez-Rodriguez, "Image quality and position variability assessment in minutiae-based fingerprint verification," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 150, no. 6, pp. 395–401, December 2003, special Issue on Biometrics on the Internet.
[17] M. Thieme, *ISO/IEC 19795-2 Biometric Performance Testing and Reporting: Scenario Testing*, cd2 ed., JTC1 / SC37 / Working Group 5, August 2005, http://isotc.iso.org/isotcportal.
[18] *The Color FERET Face Database*, National Institute of Standards and Technology, http://www.nist.gov/humanid/feret, March 2002.
[19] E. Tabassi, *NFIQ Compliance Test, NISTIR WXYZ*, National Institute of Standards and Technology, http://fingerprint.nist.gov/NFIQ, 2006.